Bayesian neural networks become heavier-tailed with depth

Mariia Vladimirova, Julyan Arbel, Pablo Mesejo

Bayesian neural networks become heavier-tailed with depth

Mariia Vladimirova
Univ. Grenoble Alpes, Inria,
CNRS, Grenoble INP, LJK,
38000 Grenoble, France
mariia.vladimirova@inria.fr

Julyan Arbel
Univ. Grenoble Alpes, Inria,
CNRS, Grenoble INP, LJK,
38000 Grenoble, France
julyan.arbel@inria.fr

Pablo Mesejo
DaSCI, Univ. of Granada,*
18071 Granada, Spain
pmesejo@decsai.ugr.es

Abstract

We investigate deep Bayesian neural networks with Gaussian priors on the weights and ReLU-like nonlinearities, shedding light on novel distribution properties at the level of the neural network units. The main thrust of the paper is to establish that the prior distribution induced on the units before and after activation becomes increasingly heavier-tailed with depth. We show that first layer units are Gaussian, second layer units are sub-Exponential, and we introduce sub-Weibull distributions to characterize the deeper layers units. This result provides new theoretical insight on deep Bayesian neural networks, underpinning their practical potential. The workshop paper is based on the original paper Vladimirova et al. (2018).

1 Introduction

Bayesian approaches investigate models by assuming a prior distribution on their parameters. Bayesian machine learning refers to extending standard machine learning approaches with posterior inference, a line of research pioneered by the works Neal (1992); MacKay (1992) on Bayesian neural networks. The interest of the Bayesian approach to NNs is at least twofold. First, it offers a principled approach for modeling uncertainty of the training procedure, which is a limitation of standard NNs that only provide point estimates. A second main asset of Bayesian models is that they represent regularized versions of their classical counterparts. For instance, maximum a posteriori (MAP) estimation of a Bayesian regression model with double exponential (Laplace) prior is equivalent to Lasso regression (Tibshirani, 1996), while a Gaussian prior leads to ridge regression. When it comes to neural networks, the regularization mechanism is also appreciated in the literature, since neural networks traditionally suffer from overparameterization, resulting in overfitting.

Central in the field of regularization techniques is the *weight decay* penalty (Krogh and Hertz, 1991), which is equivalent to MAP estimation of a Bayesian neural network with independent Gaussian priors on the weights. Srivastava et al. (2014) have suggested *dropout* as a regularization method in which neurons are randomly turned off. Gal and Ghahramani (2016) proved that the neural network trained with *dropout* is equivalent to a probabilistic model, i.e. a deep Gaussian process (Damianou and Lawrence, 2013). It leads to the consideration of such neural networks as Bayesian models.

This study is devoted to the investigation of hidden units prior distributions in Bayesian neural networks under assumption of independent Gaussian weights. We first describe a fully connected neural network architecture as illustrated in Figure 2, Appendix D. Given an input $\boldsymbol{x} \in \mathbb{R}^N$, the ℓ-th hidden layer unit activations are defined as

$$\boldsymbol{g}^{(\ell)}(\boldsymbol{x}) = \boldsymbol{W}^{(\ell)}\boldsymbol{h}^{(\ell-1)}(\boldsymbol{x}), \quad \boldsymbol{h}^{(\ell)}(\boldsymbol{x}) = \phi(\boldsymbol{g}^{(\ell)}(\boldsymbol{x})), \tag{1}$$

*Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada, Spain

Third workshop on Bayesian Deep Learning (NIPS 2018), Montréal, Canada.

where $\boldsymbol{W}^{(\ell)}$ is a weight matrix including the bias vector. A nonlinear activation function $\phi : \mathbb{R} \to \mathbb{R}$ is applied element-wise, which is called nonlinearity, $\boldsymbol{g}^{(\ell)} = \boldsymbol{g}^{(\ell)}(\boldsymbol{x})$ is a vector of pre-nonlinearities, and $\boldsymbol{h}^{(\ell)} = \boldsymbol{h}^{(\ell)}(\boldsymbol{x})$ is a vector of post-nonlinearities. When we refer to either pre- or post-nonlinearities, we will use the notation $\boldsymbol{U}^{(\ell)}$.

2 Heavy-tailed result

Definition 2.1 (Sub-Weibull random variable)**.** *A random variable X, that satisfies*

$$\mathbb{P}(|X| \geq x) \leq \exp\left(-x^{1/\theta}/K\right) \quad \textit{for all } x \geq 0.$$

for $K > 0$, is called a sub-Weibull random variable with the tail parameter $\theta > 0$, which is denoted by $X \sim \mathsf{subW}(\theta)$.

Informally, the tails of a $\mathsf{subW}(\theta)$ distribution are dominated by (i.e. decay at least as fast as) the tails of a Weibull variable with the shape parameter equal to $1/\theta$ (Rinne, 2008). The larger tail parameter θ, the heavier the tails of the sub-Weibull distribution. We refer to Vladimirova et al. (2018) for more details about sub-Weibull distribution.

2.1 Assumptions on neural network

Let all weights (including biases) be independent and have zero-mean normal distribution

$$W_{i,j}^{(\ell)} \sim \mathcal{N}(0, \sigma_w^2), \tag{2}$$

for all $1 \leq \ell \leq L$, $1 \leq i \leq H_{\ell-1}$ and $1 \leq j \leq H_\ell$. Given some input $\boldsymbol{x}$, such prior distribution induces by forward propagation (1) a prior distribution on the pre-nonlinearities and post-nonlinearities, whose *tail properties* are the focus of this section. To this aim, the nonlinearity ϕ is required to span at least half of the real line as follows. We introduce an extended version of the nonlinearity assumption from Matthews et al. (2018):

Definition 2.2 (Extended envelope property for nonlinearities)**.** *A nonlinearity $\phi : \mathbb{R} \to \mathbb{R}$ is said to obey the extended envelope property if there exist $c_1, c_2, d_1, d_2 \geq 0$ such that the following inequalities hold*

$$\begin{aligned} |\phi(u)| &\geq c_1 + d_1|u| \quad \textit{for all } u \in \mathbb{R}_+ \textit{ or } u \in \mathbb{R}_-, \\ |\phi(u)| &\leq c_2 + d_2|u| \quad \textit{for all } u \in \mathbb{R}. \end{aligned} \tag{3}$$

The interpretation of this property is that ϕ must shoot to infinity at least in one direction ($\mathbb{R}_+$ or $\mathbb{R}_-$) at least linearly (first line of (3)), and also at most linearly (second line of (3)). Compactly supported nonlinearities such as sigmoid and tanh do not satisfy the extended envelope property but the majority of other nonlinearities do, including ReLU, ELU, SELU (see Klambauer et al. (2017) for details), and others.

Lemma 2.1. *Let a nonlinearity $\phi : \mathbb{R} \to \mathbb{R}$ obey the extended envelope property. Then for any symmetric random variable X the following asymptotic equivalence[3] holds*

$$\mathbb{E}[\phi(X)^k] \asymp \mathbb{E}[X^k], \quad \textit{for } k \to \infty. \tag{4}$$

2.2 Main theorem

This section postulates the rigorous result with a proof sketch. In Supplementary material one can find proofs of intermediate lemmas and a covariance theorem which states the non-negative covariance between post-nonlinearities.

Theorem 2.1 (Sub-Weibull units)**.** *Consider a feed-forward Bayesian neural network with Gaussian priors (2) with nonlinearity ϕ satisfying the extended envelope condition of Definition 2.2. Then conditional on the input $\boldsymbol{x}$, the marginal prior distribution induced by forward propagation (1) on*

[3] See Definition B.1 for the asymptotic equivalence $\asymp$ definition in Supplementary material.

any unit (pre- or post-nonlinearity) of the ℓ-th hidden layer is sub-Weibull with optimal tail parameter $\theta = \ell/2$. That is for any $1 \leq \ell \leq L$, and for any $1 \leq m \leq H_\ell$,

$$U_m^{(\ell)} \sim \mathsf{subW}(\ell/2),$$

where a subW *distribution is defined in Definition 2.1, and $U_m^{(\ell)}$ is either a pre-nonlinearity $g_m^{(\ell)}$ or a post-nonlinearity $h_m^{(\ell)}$.*

Before proving the result, we illustrate it in Figure 1 which represents the first three hidden layers pre-nonlinearity marginal distributions (left panel). These densities are obtained as kernel density estimators from a sample of size 10^5 from the prior on the pre-nonlinearities, which is itself obtained by sampling 10^5 sets of weights $\boldsymbol{W}$ from the Gaussian prior (2) and forward propagation via (1). The three hidden layers of neural network have $H_1 = 25$, $H_2 = 24$ and $H_3 = 4$ hidden units, respectively. Being a linear combination involving symmetric weights $\boldsymbol{W}$, pre-nonlinearities $\boldsymbol{g}$ are also symmetric, thus we visualize only their positive part. The input vector $\boldsymbol{x} \in \mathbb{R}^{50}$ is sampled from a standard normal distribution once for all at the start. The nonlinearity ϕ is the ReLU function. The prior distribution of post-nonlinearities has a Dirac mass at zero with a coefficient of $1/2$ and they are no more symmetric. But the post-nonlinearity prior distribution tails remain the same as of pre-nonlinearities on $\mathbb{R}_+$, and is represented on the right panel of Figure 1.

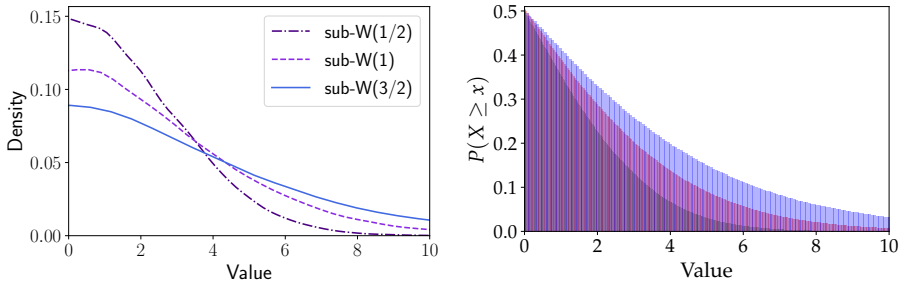

Figure 1: Illustration of the first three layers hidden units marginal prior distributions.

Proof of Theorem 2.1. The idea is to prove by induction with respect to hidden layer depth ℓ that pre- and post-nonlinearities satisfy the asymptotic moment equivalence $\|g^{(\ell)}\|_k \asymp k^{\ell/2}$ and $\|h^{(\ell)}\|_k \asymp k^{\ell/2}$. The statement of the theorem then follows by the moment characterization of optimal sub-Weibull tail coefficient in Proposition A.1.

Base step: the first hidden layer pre-nonlinearity g follows normal distribution, since

$$g = \boldsymbol{W}_i^\intercal \boldsymbol{x} \sim \mathcal{N}(0, \sigma_w^2 \|\boldsymbol{x}\|^2).$$

Then, for normal zero-mean variable g from Lemma B.1: $\|g\|_k \asymp \sqrt{k}$. As activation function ϕ obeys extended envelope property (Definition 2.2), according to Lemma B.2, nonlinearity moments are asymptotic equivalent to symmetric variable moments

$$\|\phi(g)\|_k \asymp \|g\|_k \sim \sqrt{k}.$$

It implies that first hidden layer post-nonlinearities h is sub-Gaussian or sub-Weibull with tail parameter $\theta = 1/2$ (Definition 2.1).

Inductive step: let us suppose the post-nonlinearity of $(\ell - 1)$-th hidden layer satisfies the moment condition. Hidden units satisfy the non-negative covariance theorem (Theorem C.1):

$$\mathrm{Cov}\left[\left(h^{(\ell-1)}\right)^s, \left(\tilde{h}^{(\ell-1)}\right)^t\right] \geq 0, \text{ for any } s, t \in \mathbb{N}.$$

Let the number of hidden units in $(\ell - 1)$-th layer equals to H. Then according to Lemma B.3, under assumption of zero-mean Gaussian weights, pre-nonlinearities of ℓ-th hidden layer $g^{(\ell)} = \sum_{j=1}^{H} W_{i,j}^{(\ell-1)} h_j^{(\ell-1)}$ also satisfy the moment condition, but with $\theta = \ell/2$

$$\|g^{(\ell)}\|_k \asymp k^{\ell/2}.$$

Using Lemma B.2 from Supplementary material, one can show that post-nonlinearities $h^{(\ell)}$ satisfy the same moment condition as pre-nonlinearities $g^{(\ell)}$. This finishes the proof. □

References

Damianou, A. and Lawrence, N. (2013). Deep Gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059.

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. (2017). Self-normalizing neural networks. In *Advances in Neural Information Processing Systems*, pages 971–980.

Krogh, A. and Hertz, J. A. (1991). A simple weight decay can improve generalization. In *Neural Information Processing Systems*, pages 950–957.

MacKay, D. J. (1992). A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472.

Matthews, A. G. d. G., Rowland, M., Hron, J., Turner, R. E., and Ghahramani, Z. (2018). Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*.

Neal, R. M. (1992). Bayesian training of backpropagation networks by the hybrid Monte Carlo method. Technical report, Citeseer.

Rinne, H. (2008). *The Weibull distribution: a handbook*. Chapman and Hall/CRC.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Vladimirova, M., Arbel, J., and Mesejo, P. (2018). Bayesian neural networks increasingly sparsify their units with depth. *arXiv preprint arXiv:1810.05193*.

A Equivalent sub-Weibull distribution properties

Proposition A.1 (Sub-Weibull distribution)**.** *Let X be a random variable. Then the following properties are equivalent; the parameters $K_i > 0$ appearing in these properties differ from each other by at most an absolute constant factor.*

1. *The tails of X satisfy*

$$\mathbb{P}(|X| \geq x) \leq 2\exp\left(-x^{1/\theta}/K_1\right) \quad \text{for all } x \geq 0.$$

2. *The moments of X satisfy*

$$\|X\|_k = \left(\mathbb{E}[|X|^k]\right)^{1/k} \leq K_2 k^\theta \quad \text{for all } k \geq 1.$$

3. *The MGF of $X^{1/\theta}$ satisfies*

$$\mathbb{E}\left[\exp\left(\lambda^{1/\theta} X^{1/\theta}\right)\right] \leq K_2 \exp(K_3^{1/\theta}\lambda^{1/\theta})$$

   for all λ such that $|\lambda| \leq \frac{1}{K_3}$.

4. *The MGF of $X^{1/\theta}$ is bounded at some point, namely*

$$\mathbb{E}\left[\exp\left(X^{1/\theta}/K_4\right)\right] \leq 2.$$

Proof. **1 ⇒ 2.** Assume property **1** holds. Applying the integral identity for $|X|^k$, we obtain

$$\begin{aligned}
\mathbb{E}\left[|X|^k\right] &= \int_0^\infty \mathbb{P}\left(|X|^k > x\right) dx \\
&= \int_0^\infty \mathbb{P}\left(|X| > x^{1/k}\right) dx \\
&\le \int_0^\infty 2\exp\left(-x^{1/(k\theta)}/K_1\right) dx \\
&= 2K_1^{k\theta} k\theta \int_0^\infty e^{-u} u^{k\theta-1} du = 2K_1^{k\theta} k\theta\, \Gamma(k\theta) \\
&\sim K_1^{k\theta} k\theta\, (k\theta-1)^{k\theta-1} \sim (K_1 k\theta)^{k\theta}.
\end{aligned}$$

Taking the k-th root of the expression above yields property **2**

$$\|X\|_k \lesssim (K_1\theta)^\theta k^\theta \le K_2 k^\theta,$$

with $K_2 = (K_1\theta)^\theta$.

2 ⇒ 3. Assume property **2** holds. Recalling the Taylor series expansion of the exponential function, we obtain

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda^{1/\theta} X^{1/\theta}\right)\right] &= \mathbb{E}\left[1 + \sum_{k=1}^\infty \frac{(\lambda^{1/\theta}|X|^{1/\theta})^k}{k!}\right] \\
&= 1 + \sum_{k=1}^\infty \frac{\lambda^{k/\theta}\mathbb{E}[|X|^{k/\theta}]}{k!}.
\end{aligned}$$

Property **2** guarantees that $\mathbb{E}[|X|^k] \le K_2 k^{k/\theta}$ and $\mathbb{E}[|X|^{k/\theta}] \le K_2 (k/\theta)^k$ for some K_2. Stirling's approximation yields $k! \ge (k/e)^k$. Substituting these two bounds, we get

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda^{1/\theta} X^{1/\theta}\right)\right] &\le \sum_{k=1}^\infty \frac{\lambda^{k/\theta} K_2 (k/\theta)^k}{(k/e)^k} \\
&= \sum_{k=0}^\infty K_2 (e\lambda^{1/\theta}/\theta)^k = \frac{K_2}{1 - e\lambda^{1/\theta}/\theta},
\end{aligned}$$

provided that $e\lambda^{1/\theta}/\theta < 1$, in which case the geometric series above converges. To bound this quantity further, we can use the numeric inequality $\frac{1}{1-x} \le e^{2x}$, which is valid for $x \in [0, 1/2]$. It follows that

$$\mathbb{E}\left[\exp\left(\lambda^{1/\theta} X^{1/\theta}\right)\right] \le K_2 \exp\left(2e\lambda^{1/\theta}/\theta\right)$$

for all λ satisfying $|\lambda| \le \left(\frac{\theta}{2e}\right)^\theta$. This yields property **3** with $K_3 = (2e/\theta)^\theta$.

3 ⇒ 4. Assume property **3** holds. Take $\lambda = 1/K_4$, where $K_4 \ge K_3/(\ln 2 - \ln K_2)^\theta$. This yields property **4**.

4 ⇒ 1. Assume property **4** holds. We may assume that $K_4 = 1$. Then, by Markov's inequality and property **3**, we obtain

$$\begin{aligned}
\mathbb{P}\big(|X| > x\big) &= \mathbb{P}\big(e^{|X|^{1/\theta}} > e^{x^{1/\theta}}\big) \\
&\le \frac{\mathbb{E}[e^{|X|^{1/\theta}}]}{e^{x^{1/\theta}}} \le 2e^{-x^{1/\theta}/K_1}.
\end{aligned}$$

This proves property **1** with $K_1 = 1$. □

Remark A.1. The constant 2 that appears in some properties in Proposition A.1 does not have any special meaning. It is chosen for simplicity and can be replaced by other absolute constants.

B Intermediate lemmas

Introduce the definition of asymptotic equivalence between numeric sequences:

Definition B.1 (Asymptotic equivalence)**.** *Two sequences a_k and b_k are called asymptotic equivalent and denoted as $a_k \asymp b_k$ if there exist constants $d > 0$ and $D > 0$ such that*

$$d \leq \frac{a_k}{b_k} \leq D, \quad \textit{for all } k \in \mathbb{N}. \tag{5}$$

Lemma B.1 (Gaussian moments)**.** *Let X be a normal random variable such that $X \sim \mathcal{N}(0, \sigma^2)$, then the following asymptotic equivalence holds*

$$\|X\|_k \asymp \sqrt{k}.$$

Proof. The moments of central normal absolute random variable $|X|$ are equal to

$$\mathbb{E}[|X|^k] = \int_{\mathbb{R}} |x|^k \, p(x) \, dx = 2 \int_0^{\infty} x^k \, p(x) \, dx = \frac{1}{\sqrt{\pi}} \sigma^k 2^{k/2} \Gamma\left(\frac{k+1}{2}\right). \tag{6}$$

We have the expression for the Gamma function

$$\Gamma(z) = \sqrt{\frac{2\pi}{z}} \left(\frac{z}{e}\right)^z \left(1 + \frac{1}{12z} + o\left(\frac{1}{z}\right)\right). \tag{7}$$

Substituting (7) into the central normal absolute moment (6), we obtain

$$\begin{aligned}
\mathbb{E}[|X|^k] &= \frac{1}{\sqrt{\pi}} \sigma^k 2^{k/2} \sqrt{\frac{4\pi}{k+1}} \left(\frac{k+1}{2e}\right)^{(k+1)/2} \left(1 + \frac{1}{6(k+1)} + o\left(\frac{1}{k}\right)\right) \\
&= \frac{2\sigma^k}{\sqrt{2e}} \left(\frac{k+1}{e}\right)^{k/2} \left(1 + \frac{1}{6(k+1)} + o\left(\frac{1}{k}\right)\right).
\end{aligned}$$

Then the roots of absolute moments can be written in the form of

$$\begin{aligned}
\|X\|_k &= \frac{\sigma}{e^{1/(2k)}} \sqrt{\frac{k+1}{e}} \left(1 + \frac{1}{6(k+1)} + o\left(\frac{1}{k}\right)\right)^{1/k} \\
&= \frac{\sigma}{e} \frac{\sqrt{k+1}}{e^{1/(2k)}} \left(1 + \frac{1}{6(k+1)k} + o\left(\frac{1}{k^2}\right)\right) \\
&= \frac{\sigma}{e} c_k \sqrt{k+1}.
\end{aligned}$$

Here the coefficient c_k denotes

$$c_k = \frac{1}{e^{1/(2k)}} \left(1 + \frac{1}{6(k+1)k} + o\left(\frac{1}{k^2}\right)\right) \to 1,$$

with $k \to \infty$. Thus, asymptotic equivalence holds

$$\|X\|_k \asymp \sqrt{k+1} \asymp \sqrt{k}.$$

□

Lemma B.2 (Nonlinearity moments)**.** *Let X be symmetric random variable, $\phi(x)$ be nonlinear function satisfying extended envelope property, then the following asymptotic equivalence holds*

$$\|\phi(X)\|_k \asymp \|X\|_k.$$

Proof. According to extended envelope property, $\mathbb{E}[\phi(X)^k] \asymp \mathbb{E}[X^k]$. That means there exist constants d and D such that for all $k \in \mathbb{N}$ it holds

$$d \leq \frac{\mathbb{E}[\phi(X)^k]}{\mathbb{E}[|X|^k]} \leq D.$$

Observing that

$$d' \leq d^{1/k} \leq \frac{\|\phi(X)\|_k}{\|X\|_k} \leq D^{1/k} \leq D',$$

the bounding constants are $d' = \min\{1, d\}$, $D' = \max\{1, D\}$. It yields asymptotic equivalence

$$\|\phi(X)\|_k \asymp \|X\|_k.$$

The lemma is proved. □

Lemma B.3 (Multiplication moments)**.** *Let W and X be independent random variables such that $W \sim \mathcal{N}(0, \sigma^2)$ and for some $p > 0$ it holds*

$$\|X\|_k \asymp k^p. \tag{8}$$

Let W_i be independent copies of W, and X_i be copies of X, $i = 1, \dots, H$ with non-negative covariance between moments of copies

$$Cov\left[X_i^s, X_j^t\right] \geq 0, \quad \text{for } i \neq j,\ s, t \in \mathbb{N}. \tag{9}$$

Then we have the following asymptotic equivalence

$$\left\| \sum_{i=1}^{H} W_i X_i \right\|_k \asymp k^{p+1/2}. \tag{10}$$

The statement proof is based on mathematical induction, see Vladimirova et al. (2018).

C Covariance theorem

Theorem C.1 (Non-negative covariance between hidden units)**.** *Consider the deep neural network described in Section 3 with assumptions from Section 3.1. The covariance between hidden units of the same layer is non-negative. Moreover, for given ℓ-th hidden layer units $h^{(\ell)}$ and $\tilde{h}^{(\ell)}$, it holds*

$$Cov\left[\left(h^{(\ell)}\right)^s \left(\tilde{h}^{(\ell)}\right)^t\right] \geq 0, \ \text{where } s, t \in \mathbb{N}.$$

For first hidden layer $\ell = 1$ there is equality for all s and t.

We refer to Vladimirova et al. (2018) for the proof.

D Illustration

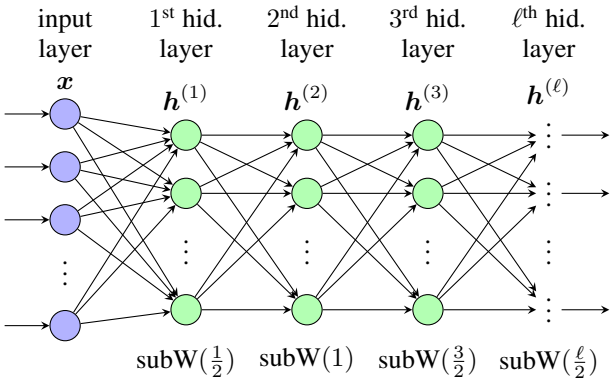

Figure 2: Neural network architecture and characterization of the ℓ-layer units prior distribution as sub-Weibull distribution with tail parameter $\ell/2$ (see Definition 2.1).